

# Revisiting Neural Networks for Continual Learning: An Architectural Perspective

Aojun Lu<sup>1</sup>, Tao Feng<sup>2</sup>, Hangjie Yuan<sup>3</sup>, Xiaotian Song<sup>1</sup> and Yanan Sun<sup>1\*</sup>

<sup>1</sup>Sichuan University

<sup>2</sup>Tsinghua University

<sup>3</sup>Zhejiang University

aojunlu@stu.scu.edu.cn, fengtao.hi@gmail.com, hj.yuan@zju.edu.cn  
songxt@stu.scu.edu.cn, ysun@scu.edu.cn

## Abstract

Efforts to overcome catastrophic forgetting have primarily centered around developing more effective Continual Learning (CL) methods. In contrast, less attention was devoted to analyzing the role of network architecture design (e.g., network depth, width, and components) in contributing to CL. This paper seeks to bridge this gap between network architecture design and CL, and to present a holistic study on the impact of network architectures on CL. This work considers architecture design at the network scaling level, i.e., width and depth, and also at the network components, i.e., skip connections, global pooling layers, and down-sampling. In both cases, we first derive insights through systematically exploring how architectural designs affect CL. Then, grounded in these insights, we craft a specialized search space for CL and further propose a simple yet effective **ArchCraft** method to steer a CL-friendly architecture, namely, this method recrafts *AlexNet/ResNet* into *AlexAC/ResAC*. Experimental validation across various CL settings and scenarios demonstrates that improved architectures are parameter-efficient, achieving state-of-the-art performance of CL while being **86%**, **61%**, and **97%** more compact in terms of parameters than the naive CL architecture in *Task IL* and *Class IL*. Code is available at <https://github.com/byyx666/ArchCraft>.

## 1 Introduction

Artificial Intelligence (AI) is expected to feature a robust capability to continuously acquire and update knowledge like creatures. To achieve this, Continual Learning (CL), also known as Incremental Learning (IL), has garnered much attention in AI [Wang *et al.*, 2023]. Early, some efforts [Li and Hoiem, 2017; Serra *et al.*, 2018] centered on Task Incremental Learning (*Task IL*), where the task identity is given during both training and inference phases to assign a classifier. Recently, more works [Feng *et al.*, 2022; Chen and Chang, 2023; Bian *et al.*, 2024] focused on the more intricate Class Incre-

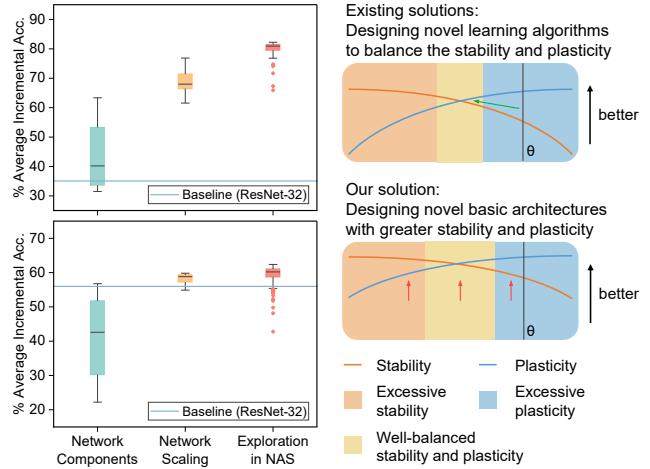


Figure 1: **(Left)** Impact of architectural designs on CL in *Task IL* (top) and *Class IL* (bottom). This displays the distributions of the CL performance with variations in network components and scaling. **(Right)** Illustration of **ArchCraft** for improving CL performance. Existing methods aim to seek the optimal balance of stability and plasticity (green arrow), ArchCraft focuses on enhancing stability and plasticity by recrafting the basic architecture (red arrows).

mental Learning (*Class IL*), in which the task identity is accessible only during training.

In both *Task IL* and *Class IL*, it is essential for the model to possess the plasticity required to acquire new knowledge and the stability necessary to retain previously learned knowledge. However, the stability-plasticity dilemma remains a persistent challenge in CL [Grossberg, 2013]. To address this, existing works [Kirkpatrick *et al.*, 2017; Rebuffi *et al.*, 2017; Shibata *et al.*, 2021; Wang *et al.*, 2022] focus on developing novel algorithms to trade-off stability and plasticity. Among them, expansion-based methods dedicate different incremental network architectures to minimize conflicts between stability and plasticity. For instance, DER [Yan *et al.*, 2021] centers on feature extractor scaling, by contrast, MEMO [Zhou *et al.*, 2022] puts effort into specific layers. In essence, these efforts resort to network width to build a good CL model. Moreover, some works can be formulated as utilizing network components to resist forgetting, e.g., CN [Pham *et al.*, 2022], FPF [Zhao *et al.*, 2023], and LoRA [Lin *et al.*, 2021]. However, these methods are built upon existing network ar-

\*Corresponding author.

chitectures, which limits the scope of architecture exploration for better stability and plasticity.

A clear trend is that expanded architectures outperform the original ones. This raises a concern about whether the existing basic architectures of the network are ideal for CL. As shown in Figure 1, we observe that (i) various architectural designs significantly shape the impact towards CL, and (ii) directly using the native ResNet [He *et al.*, 2016] leads to performance far from satisfactory. This motivates us to design architectures tailored for CL.

#### How to seek an efficient and CL-friendly architecture?

Naturally, Neural Architecture Search (NAS) [Elsken *et al.*, 2019] is a straightforward candidate for exploring architectures for CL due to the automated search that can be performed on architectures. In this paper, we employ NAS as a promising solution for CL and probe catastrophic forgetting from an architectural perspective. However, the performance of NAS hinges on the quality of the search space design [Radosavovic *et al.*, 2020]. Given the limited prior knowledge regarding the influence of architectural designs on the performance of CL, constructing a suitable search space remains challenging. Thus, it becomes imperative to empirically glean this prior knowledge before employing NAS.

In short, this paper aims to bridge the gap between network architectures and CL. To this end, we (i) systematically scrutinize the influence of architectural design on CL, (ii) identify critical design candidates that can boost CL, and (iii) propose a NAS method with a novel search space tailored for CL.

Given that ResNet stands as a cornerstone in the AI era, we initiate this work from it to revisit neural networks and assess the impact of architectural designs on CL. Consequently, we formulated this study around network scaling and network components. Concerning the former, we consider the network width and depth. Regarding the latter, we delve into the investigation of skip connections, global pooling layers, and down-sampling approaches. Subsequently, grounded in empirical observations, we craft a specialized search space for CL, each candidate architecture within this space is succinctly represented, facilitating an efficient search process. In conclusion, we propose a simple yet effective method to **Craft** CL-friendly **Architectures**, dubbed **ArchCraft** (*AC*). ArchCraft recrafts *AlexNet/ResNet* into *AlexAC/ResAC* to guide a well-designed network architecture for CL. As a result, ArchCraft achieves superior CL performance while employing significantly fewer parameters than the original architecture.

To sum up, the contributions of this work are as follows:

- We thoroughly scrutinize the mechanisms through which neural architectural design affects CL, and demonstrate that wider and shallower architectures better align with an effective CL model.
- To the best of our knowledge, this paper is the pioneering effort to employ neural architecture design to shape a CL-friendly architecture. To achieve this, we propose a novel ArchCraft method to steer an architecture with greater stability and plasticity.
- Extensive experiments show that the proposed method is parameter-efficient across various CL scenarios, i.e., achieving state-of-the-art performance with much fewer

parameters than the baseline. Furthermore, our method features a controllable parameter.

## 2 Related work

### 2.1 Continual Learning and Neural Architectures

**Continual Learning.** Neural networks have achieved remarkable success in various fields [Yuan *et al.*, 2023; Zhu *et al.*, 2023], but are ill-equipped for CL due to catastrophic forgetting. A series of works, such as encompassing regularization, dedicated memory systems, and modular architectures, are proposed to overcome forgetting. In detail, Regularization-based methods regularize the variation of network parameters or logit outputs to maintain stability, e.g., EWC [Kirkpatrick *et al.*, 2017] and XK-FAC [Lee *et al.*, 2020]. Replay-based methods preserve a few exemplars from previous tasks and replay them in learning a new task to prevent forgetting, e.g., iCaRL [Rebuffi *et al.*, 2017] and Gcr [Tiwari *et al.*, 2022]. This category of work designs different sampling strategies to establish the limited memory buffer for replay. Expansion-based methods continually expand the network methods, in which different incremental parts of the networks are dedicated to specific tasks, e.g., BNS [Qin *et al.*, 2021] and MEMO [Zhou *et al.*, 2022]. This category of work focuses on the efficient utilization of network architectures.

**Neural Architectures for CL.** In principle, the performance of neural networks is mainly decided by their weights and architectures. The CL methods mentioned above are mainly for the weights, while the research on neural architectures for enhancing CL is still in the infant. In the literature, the work [Mirzadeh *et al.*, 2022a] concluded that wider network architectures suffer less catastrophic forgetting in *Task IL*. Upon this, another work [Mirzadeh *et al.*, 2022b] further investigated the impact of certain network components on *Task IL*. To the best of our knowledge, there is no systematic work dedicated to designing the CL-friendly basic architectures (i.e., the backbones).

To sum up, unlike these methods that focus on expanding or modular the specific architecture to improve CL, this work centers on the architecture itself. Notably, we extend the scope of architectural design to encompass a broader scenario, a.k.a more intricate *Class IL*. And we further propose an effective method for designing enhanced network architectures tailored for CL.

### 2.2 Neural Architecture Search (NAS)

NAS automates the design of high-performance neural architectures by formulating the design process as an optimization problem [Elsken *et al.*, 2019]. In this process, NAS employs an optimization method (i.e., search strategy), to traverse a predefined search space comprising candidate architectures. After searching, the architecture demonstrating the best performance is chosen as the final design. Some works have already empirically shown that NAS can craft architectures that surpass those manually designed [Zoph *et al.*, 2018].

**Discussion.** NAS can be employed to design architectures for CL. However, the crux of NAS lies in defining a suitable search space [Radosavovic *et al.*, 2020; Wan *et al.*, 2022], which delineates a searchable subset of potential architectures

Network Components			Performance in Task IL				Performance in Class IL			
Down.	Skip	GAP	R32-LA	R32-AIA	R18-LA	R18-AIA	R32-LA	R32-AIA	R18-LA	R18-AIA
Strided	✓	✓	25.80±1.80	35.06±0.88	38.12±2.85	49.68±0.91	37.92±0.64	55.97±0.85	40.34±0.95	57.79±1.09
	✓	✗	55.52±1.41	60.89±1.83	62.95±3.45	65.41±1.92	22.63±3.36	27.37±11.18	38.85±1.60	50.29±3.01
	✗	✓	25.29±1.48	31.74±0.90	30.33±2.28	41.75±0.89	30.74±2.11	46.99±1.48	38.45±0.66	56.86±0.82
	✗	✗	38.30±1.17	45.58±1.76	57.66±1.72	62.13±1.02	27.63±2.79	38.66±6.24	33.99±3.77	45.98±4.53
Max Pooling	✓	✓	25.91±2.35	35.58±0.78	39.89±1.80	53.15±0.69	<b>38.27</b> ±0.88	<b>56.79</b> ±0.88	<b>40.50</b> ±0.34	<b>59.53</b> ±1.26
	✓	✗	<b>57.31</b> ±1.56	<b>63.35</b> ±1.65	<b>63.94</b> ±3.13	<b>68.74</b> ±3.27	24.91±3.58	22.23±13.19	40.00±0.67	54.11±1.79
	✗	✓	24.54±0.53	32.16±0.87	30.46±2.52	42.24±1.08	30.69±1.27	47.50±1.42	37.34±0.72	56.53±0.83
	✗	✗	36.53±0.85	44.72±1.31	60.16±0.88	65.41±0.69	16.67±13.00	25.64±18.06	33.92±3.98	47.83±4.96
Avg Pooling	✓	✓	24.37±1.72	35.00±0.92	38.47±2.01	53.52±0.67	38.25±0.52	56.67±0.61	39.85±1.52	57.53±1.14
	✓	✗	56.16±2.93	62.60±1.74	63.92±3.22	68.56±2.78	27.62±5.27	34.69±11.64	37.35±3.96	49.74±4.24
	✗	✓	23.70±1.92	31.50±0.88	30.21±1.27	42.82±0.80	30.33±2.09	46.55±1.24	37.02±0.92	55.99±0.48
	✗	✗	35.86±2.09	44.93±2.42	60.68±0.86	65.78±1.29	23.04±11.10	33.05±14.94	34.84±1.67	47.63±1.41

Table 1: The CL performance of the networks with different configurations of down-sampling approaches (denoted as ‘Down.’) and whether to use skip connections (denoted as ‘Skip’) and GAP or not. ‘R18’ and ‘R32’ represent networks based on ResNet-18 and ResNet-32.

from the vast architecture space. While numerous architectural design practices offer substantial prior knowledge about architectures with remarkable plasticity for standard learning paradigms [Howard *et al.*, 2019; Tan and Le, 2021], designing a search space specifically tailored for CL requires revisiting existing architectural design experiences. This is essential as architectures for CL need to possess not only plasticity but also stability [Grossberg, 2013], necessitating the construction of a CL-friendly search space.

### 3 Preliminaries

In this section, we describe the experimental setup for training and evaluating the networks toward CL.

**Benchmark.** For the CL scenarios mentioned above, i.e., *Task IL* and *Class IL*, we assess network performance on CIFAR-100. The training process involves gradually introducing all 100 classes with 5 classes per incremental step, totaling 20 steps or tasks. In *Task IL*, the network is trained using a vanilla SGD optimizer, while in *Class IL*, a replay buffer containing 2,000 examples is employed.

**Evaluation Metrics.** Let  $K$  be the number of tasks, the average classification accuracy after learning the  $b$ -th task, say  $A_b$ , is defined as:

$$A_b = \frac{1}{b} \sum_{i=1}^b a_{i,b} \quad (1)$$

where  $a_{i,b}$  is the classification accuracy evaluated on the test set of the  $i$ -th task after learning the  $b$ -th task ( $i \leq b$ ). The performance of CL is measured by two metrics: the *Last Accuracy* (LA) and the *Average Incremental Accuracy* (AIA). The LA is the classification accuracy after the last task, i.e.,  $LA = A_K$ , which reflects the overall accuracy among all classes. Further, the AIA denotes the average of  $A_b$  over all tasks, i.e.,  $AIA = \frac{1}{K} \sum_{b=1}^K A_b$ , which reflects the performance of all incremental stages. The higher LA and AIA, the better CL performance.

**Implementation Details.** In all experiments, we report the mean and standard deviation calculated across 5 runs with different random seeds. For *Task IL*, we train the model by 60 epochs in the first task and 20 epochs in the subsequent

tasks. For *Class IL*, we follow PyCIL [Zhou *et al.*, 2023] to train the model by 200 epochs in the first task and 70 epochs in the subsequent tasks.

## 4 Design of CL-Friendly Architectures

In this section, we first decompose and study the architectural design on the performance of *Task IL* and *Class IL* at the network components and scaling. Upon these insights, we propose an ArchCraft method to craft CL-friendly architectures.

### 4.1 Impact of Network Components

Strided convolution, skip connections, and Global Average Pooling (GAP) have all proven effective in neural networks in standard learning paradigms [He *et al.*, 2016]. However, the impact of these components on CL remains elusive. To address this gap, we perform experiments to determine the optimal configurations of these components for both *Task IL* and *Class IL*. To be specific, the candidate configurations can be obtained by replacing strided convolution with average/max pooling, removing skip connections, or removing GAP. In particular, we utilize the ResNet series as the architectural framework and examine the impact of the aforementioned components on CL performance.

As shown in Table 1, different configurations of network components influence the performance of both *Task IL* and *Class IL*. In brief, we observe that the optimal configuration involves employing max pooling and skip connections while removing GAP for *Task IL*. Moreover, the optimal configuration for *Class IL* is similar to *Task IL* but retains GAP. In particular, skip connections significantly improve the performance in both *Task IL* and *Class IL*, mirroring their role in standard learning paradigms. In contrast, the function of GAP diverges significantly between *Task IL* and *Class IL*. In *Task IL*, eliminating the GAP in networks considerably enhances CL performance, while in *Class IL*, it causes severe performance degradation.

### 4.2 Impact of Network Scaling

Width and depth are two crucial factors that affect the capacity and complexity of neural networks. We conduct a series

of experiments on networks with different widths and depths to investigate the impact on CL performance. Specifically, we adopt ResNet-32 as the skeleton of the network [He *et al.*, 2016], which consists of a single convolution layer and three subsequent stages, each of them containing the same number of blocks. In that case, the width is represented by the initial number of channels, denoted by  $W$ , and the depth is represented by the number of blocks in each stage, denoted by  $D$ . Moreover, we note that all of the network architectures are modified by applying the optimal configurations of network components summarized in Section 4.1.

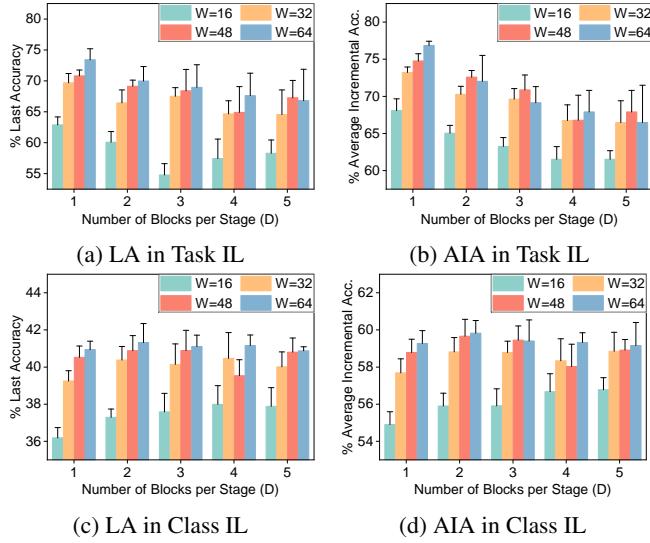


Figure 2: The performance of CL of ResNets with different network depth and width.

Figure 2 compares the CL performance of networks with different  $W$  and  $D$ . We observe that in both *Task IL* (see Figure 2a and 2b) and *Class IL* (see Figure 2c and 2d), the network with a larger width tends to exhibit better performance of CL in most cases. This phenomenon cannot be simply attributed to over-parameterization, as increasing depth does not yield similar results. In contrast, increasing the depth of the networks can even lead to obvious performance degradation in *Task IL* (see Figure 2a and 2b). These results indicate that wider and shallower networks may be more suitable for CL, explaining why ResNet-18 empirically shows better performance than ResNet-32 in Table 1. Therefore, we conclude that scaling the network depth and width appropriately is a potential way to enhance CL performance.

The above results raise a concern about whether the increase in the overall network width (i.e., initial width) or the width of the classifier (i.e., final width) contributes to the improvement of CL performance. To investigate this, we conduct further experiments on networks with the same final width but varying initial widths. In detail, we consider the network architectures with  $D = 1$ . To allow for adjustments to the final width, an additional convolutional layer with a  $1 \times 1$  kernel size is inserted before the classifier for these networks. We maintain the final width of the network with  $W = 64$  constant and adjust the output channels of the in-

serted convolutional layer for other networks to match it. The CL performance of networks with different initial and final widths are reported in Table 2.

Initial Width	Final Width	Performance in Task IL		Performance in Class IL	
		LA	AIA	LA	AIA
16	64	59.48 $\pm$ 0.03	65.65 $\pm$ 0.89	35.82 $\pm$ 0.55	54.43 $\pm$ 0.47
	256	71.51 $\pm$ 0.87	73.89 $\pm$ 0.80	36.16 $\pm$ 0.52	55.18 $\pm$ 0.63
32	128	68.24 $\pm$ 0.72	71.74 $\pm$ 0.69	39.44 $\pm$ 0.81	57.83 $\pm$ 0.72
	256	74.18 $\pm$ 0.94	76.18 $\pm$ 0.46	40.02 $\pm$ 0.74	58.02 $\pm$ 0.87
48	192	70.96 $\pm$ 1.70	74.94 $\pm$ 0.80	40.52 $\pm$ 0.92	58.93 $\pm$ 0.77
	256	73.12 $\pm$ 1.46	76.64 $\pm$ 0.63	40.53 $\pm$ 0.75	59.08 $\pm$ 0.67
64	256	73.14 $\pm$ 1.51	76.61 $\pm$ 0.66	40.97 $\pm$ 0.56	59.48 $\pm$ 0.85

Table 2: The CL performance on different width configurations.

As shown in Table 2, the results of “performance in *Task IL*” indicate that increasing the classifier width significantly benefits *Task IL*. Moreover, networks with different initial widths but the same final width show comparable performance of CL. This implies that the classifier width, rather than the overall network width, is the main factor affecting the performance of wider networks in *Task IL*. It also suggests that the adverse effect of GAP on *Task IL* may stem from the reduction in the width of the classifier. By contrast, we observe from the results of “performance in *Class IL*” that the impact of initial width on *Class IL* is as essential as classifier width. This indicates that the overall network width and the width of the classifier both significantly affect *Class IL*.

### 4.3 ArchCraft Method

To further investigate architectural designs that benefit CL, we propose ArchCraft, a straightforward yet powerful NAS method for CL. ArchCraft comprises a CL-friendly search space and employs a simple search strategy grounded in genetic algorithms.

#### Search Space towards ArchCraft

**What architectural design elements can be predetermined based on the aforementioned findings?** The above experiments show that ResNet can achieve good CL performance with slight modification, which demonstrates the great generalizability of its overall architecture for CL. Therefore, in our designed CL-friendly search space, the overall architecture of the networks is predetermined following the design of ResNet, i.e., each network architecture consists of a stem (i.e., a convolutional layer) and several units. Since the promising architecture is relatively shallow according to the experiments in Section 4.2, we use a simple unit that comprises a single convolutional layer and a skip connection to explore more flexible architecture. More specifically, each convolutional layer is equipped with a  $3 \times 3$  convolutional kernel and followed by batch normalization and ReLU. In addition, the network components are determined according to the corresponding optimal configurations summarized in Section 4.1 for *Task IL* and *Class IL*, respectively.

**What architectural design elements require further exploration?** To achieve a better CL performance, further exploration of certain architectural design elements is still necessary. First, although the significance of network width and

depth for CL has been demonstrated, determining their optimal values necessitates further investigation. Second, the positions at which the channel number is increased and feature map down-sampling occurs are critical for CL performance, as they significantly influence the feature extraction capability of the network. However, their impact on CL has not been thoroughly examined in prior experiments. Therefore, the optimal designs of the locations of these operations are still necessary to search for. In summary, the architectural variations within our proposed search space encompass network width, depth, locations for increasing the channel number, and locations for down-sampling the feature map.

**How to concisely represent the candidate network architectures in search space?** We suggest an effective scheme to encode the architectures into a concise form that facilitates searching and adaptive scaling of parameters. The number of units and the initial number of channels are each denoted by a single code, which respectively represents the network depth and width. We specify that the size of the feature map can be halved zero to five times at chosen locations, which is indicated by five codes. Specifically, if the code  $x$  is less than the number of units, it indicates that the size of the feature map is reduced by half through max pooling before the  $(x + 1)$ -th unit; otherwise, it holds no significance. Similarly, the locations of the units in which the number of output channels is doubled are indicated by the other five codes. By the above means, each architecture in the proposed search space can be simply yet concisely encoded as a sequence comprising 12 codes. Additionally, the parameters of networks can be easily scaled by simultaneously reducing the network width and depth when the number of parameters exceeds the limit.

### Search Strategy towards ArchCraft

In our evolutionary search strategy, the search process consists of an initialization phase and iterative evolutionary phases. During the initialization phase, we randomly generate a population of candidate architectures (i.e., individuals) within the defined search space. Subsequently, the fitness is evaluated for each individual. In each iteration of the evolutionary phase, the offspring is generated through the mutation after fitness evaluation. In detail, to generate an offspring individual, (i) two individuals are randomly selected from the population, and the one with higher fitness is chosen as the parent, (ii) the parent is copied and a randomly selected code of it is modified. In particular, when the code representing the number of units is modified, the codes that indicate the locations to down-sample feature maps and raise channel numbers are adjusted accordingly to keep the relative positions fixed. After the offspring generation, individuals are evaluated for fitness, and those with higher fitness are selected from the previous population and offspring to form the next population. Upon completion of the search process, the architecture exhibiting the highest fitness is chosen as the final design.

### Performance Evaluation towards ArchCraft

We employ a direct method to evaluate the performance (i.e., fitness) of searched architectures. Specifically, each architecture is trained and assessed with the settings described in Section 3, with the AIA serving as the fitness metric.

## 5 Evaluation of the Improved Architectures

We apply the proposed ArchCraft method to design improved architectures for *Task IL* and *Class IL* separately, with a population size of 10 and an evolution of 20 generations. Then, we empirically compare these architectures with the baseline ones to demonstrate the efficacy of ArchCraft.

### 5.1 Better CL Performance

#### Benchmark

We choose CIFAR-100 and Imagenet-100 to evaluate the ArchCraft-guided architectures. Both datasets are divided into 20 tasks of 5 classes each and 10 tasks of 10 classes each to construct four benchmarks: C100-inc5, C100-inc10, I100-inc5, and I100-inc10. It should be highlighted that since the ArchCraft-guided architectures are crafted based on C100-inc5, the experiments conducted on C100-inc10 and ImageNet-100 serve to validate their generalizability.

#### Evaluations in Task IL

**AlexAC.** AlexNet [Krizhevsky *et al.*, 2012] is widely used as the basic architectures in *Task IL* [Serra *et al.*, 2018; Konishi *et al.*, 2023]. Thus, we craft two architectures with distinct parameter sizes based on them: **AlexAC-A** and **AlexAC-B** featuring marginally and far fewer parameters than AlexNet.

**Setup.** For CIFAR-100, we compare ArchCraft with several classical Task CL methods: SI [Zenke *et al.*, 2017], HAT [Serra *et al.*, 2018], WSN [Kang *et al.*, 2022], and SPG [Konishi *et al.*, 2023]. To make this comparison, we train AlexAC-A and AlexAC-B using a vanilla SGD optimizer, while AlexNet is trained with various CL methods. Additionally, we report the upper-bound performance for all networks. Note that the term ‘Upper-bound’ refers to the result of training a separate model for each task, thereby eliminating any forgetting. Furthermore, we also provide a comparison between AlexNet and AlexAC-A on Imagenet-100, both of which are trained using a vanilla SGD optimizer.

	Network	#P (M)	Method	C100-inc5	C100-inc10
AlexNet	6.71	6.28 <small>↓ 6%</small>	Upper Bound	81.86	73.07
			SGD	53.78	56.92
			SI	70.3	62.9
			HAT	71.8	62.8
			SPG	75.9	67.7
			WSN	76.9	69.3
AlexAC-A	6.28 <small>↓ 6%</small>	6.28 <small>↓ 6%</small>	Upper Bound	83.59	75.09
			SGD	<b>82.38 (+5.48)</b>	<b>73.91 (+4.61)</b>
AlexAC-B	0.92 <small>↓ 86%</small>	0.92 <small>↓ 86%</small>	Upper Bound	83.58	74.72
			SGD	<b>79.32 (+2.42)</b>	<b>70.87 (+1.57)</b>

Table 3: The last accuracy of the AlexAC and AlexNet in *Task IL*. ‘#P’ represents the number of parameters of the network used. **Bolded** indicates best performance. Underline indicates second best.

Table 3 details the comparison between the AlexAC series and AlexNet on CIFAR-100. It is evident that the AlexAC series consistently outperforms AlexNet in all cases. In particular, when both are trained using the vanilla SGD, AlexAC-A achieves 28.6% and 16.99% higher LA than the AlexNet on C100-inc5 and C100-inc10 settings, respectively, while utilizing fewer parameters. The magnitude of this improvement

surpasses that achieved by state-of-the-art methods (28.6% and 16.99% vs. 23.12% and 12.38%). These results underscore the pivotal role of network architectures in CL. This implies that advancements in CL can be propelled not only through learning methods but also through improved architectural designs. Moreover, AlexAC-B attains superior performance compared to state-of-the-art methods with 86% fewer parameters. In conclusion, these observations strongly emphasize the significance of enhanced architecture in CL.

Network	#P (M)	Method	I100-inc5	I100-inc10
AlexNet	6.71	SGD	35.36	33.38
AlexAC-A	6.28 $\downarrow$ 6%	SGD	<b>52.02</b>	<b>45.86</b>

Table 4: The last accuracy on Imagenet-100 (I100) in *Task IL*.

Table 4 reports the comparison between AlexAC-A and AlexNet on Imagenet-100. We observe that AlexAC-A also achieves significantly higher LA (16.66% and 12.48%) with fewer parameters compared to AlexNet. This result demonstrates that ArchCraft can design generally useful architectures that facilitate *Task IL*.

### Evaluations in Class IL

**ResAC.** ResNet series are widely used as the basic architectures in *Class IL*. Thus, we craft two architectures with distinct parameter sizes based on them: **ResAC-A** and **ResAC-B** featuring fewer parameters than ResNet-18 and ResNet-32.

**Setup.** We evaluate our method combined with several classical Class CL methods: Replay, iCaRL [Rebuffi *et al.*, 2017], WA [Zhao *et al.*, 2020], and Foster [Wang *et al.*, 2022]. Hyper-parameters for all methods adhere to the settings in the open-source library [Zhou *et al.*, 2023], and a fixed memory size of 2,000 exemplars is utilized for all methods.

Table 5 reports the LA and AIA of the ResAC series and ResNet series in *Class IL*. In this case, ResAC-A consistently outperforms ResNet-18 across all methods, datasets, and incremental steps. In particular, ResAC-A achieves a maximum 8.19% higher LA and 8.02% higher AIA with 23% fewer parameters than the ResNet-18. Moreover, ResAC-A demonstrates superior performance on ImageNet, indicating robust generalizability. This suggests that the enhancement afforded by the improved network architecture can effectively generalize to various methods and settings. Furthermore, under stricter parameter constraints, ResAC-B, compared to ResNet-32, attains up to 5.72% and 4.98% improvements in LA and AIA with 4% fewer parameters. In conclusion, the proposed method can serve as a valuable complement to CL.

### 5.2 Reducing the Number of Parameters

In this subsection, we thoroughly examine the effects of parameter sizes and network architecture. To simplify, we choose one typical method Foster as a baseline.

First, Figure 3a yields three conclusions: (i) ArchCraft can craft neural networks with controllable parameter sizes, e.g., from ResAC-0.18M to ResAC-8.63M. (ii) ResAC series achieve better CL performance with significantly fewer parameters than the naive ResNet series, e.g., with similar performance, 61% and 97% fewer parameter sizes (from 0.46M

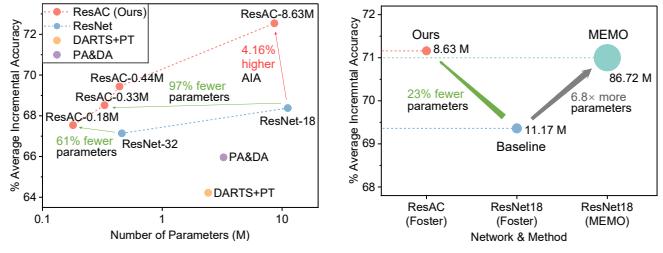


Figure 3: (a) **Performance of CL vs. Number of Parameters** on C100-inc10. (b) The comparison of ArchCraft and expansion-based method (i.e., MEMO) in terms of parameters on I100-inc10.

to 0.18M and from 11.17M to 0.33M, respectively); With similar parameter sizes, 2.3% higher performance on AIA (from 67.14% to 69.44%), or more. (iii) Existing state-of-the-art NAS methods, e.g., DARTS+PT [Wang *et al.*, 2021] and PA&DA [Lu *et al.*, 2023], fail to design CL-friendly architectures. For this point, we think that most existing NAS methods can not adaptively adjust the parameter sizes, since the key architectural elements (i.e., depth, width) that determine the number of parameters rely on predefined manually in their search space [Liu *et al.*, 2018; Wan *et al.*, 2022]. In summary, ArchCraft recrafts AlexNet/ResNet into AlexAC/ResAC to guide a well-designed network architecture for CL with fewer parameter sizes.

In addition, as Figure 3b, we compare the proposed method with the expansion-based method MEMO [Zhou *et al.*, 2022]. As is known to all, this category usually obtains performance improvement at the expense of increasing the parameter size of the baseline (from 11.17M to 86.72M). This leads to the limited availability of methods of this category. In radical contrast, ArchCraft achieves a better result with much fewer parameters than the baseline architecture (from 11.17M to 8.63M). To sum up, these empirical results provide fresh insights for advancing CL.

### 5.3 Stronger Stability and Plasticity

To further scrutinize the effectiveness of ArchCraft, we report the average forgetting and accuracy of the new task on C100-inc10 for *Task IL* and *Class IL*. As illustrated in Figure 4, the ArchCraft-guided architectures exhibit less forgetting on the previous task and higher accuracy on the current task than the baselines in a consistent setting. This observation suggests that stability and plasticity can be concurrently improved by employing architectures designed by ArchCraft. Furthermore, it emphasizes the critical role of heightened stability in enhancing CL performance.

**What contributes to the stronger stability of the ArchCraft-guided network architecture?** To investigate this, we quantitatively assess the similarity of network representations across models from all incremental stages during CL. In detail, we utilize Centered Kernel Alignment (CKA) [Kornblith *et al.*, 2019], a robust and widely used metric, to compute the similarity between representations. To facilitate this comparison, we extract feature maps from the final convolutional layer of each model by inputting the same set of instances. Subsequently, CKA is applied to measure

Method	Network	#P (M)	C100-inc5		C100-inc10		I100-inc5		I100-inc10		Max Improvement	
			LA	AIA	LA	AIA	LA	AIA	LA	AIA	LA	AIA
Replay	ResNet-32	0.46	39.10	58.17	40.02	58.21	-	-	-	-	-	-
	ResAC-B	0.44( $\downarrow$ 4%)	<b>40.45</b>	<b>59.67</b>	<b>42.79</b>	<b>59.99</b>	-	-	-	-	+2.77	+1.78
iCaRL	ResNet-18	11.17	40.04	58.80	43.23	60.42	36.30	57.30	41.00	59.21	-	-
	ResAC-A	8.63( $\downarrow$ 23%)	<b>42.99</b>	<b>62.52</b>	<b>46.62</b>	<b>63.36</b>	<b>36.78</b>	<b>57.40</b>	<b>42.44</b>	<b>60.07</b>	+3.39	+3.72
WA	ResNet-32	0.46	46.67	63.47	48.80	64.18	-	-	-	-	-	-
	ResAC-B	0.44( $\downarrow$ 4%)	<b>47.94</b>	<b>64.17</b>	<b>50.11</b>	<b>64.42</b>	-	-	-	-	+1.31	+0.70
Foster	ResNet-18	11.17	47.32	64.13	52.77	66.04	44.10	62.36	50.98	67.11	-	-
	ResAC-A	8.63( $\downarrow$ 23%)	<b>52.6</b>	<b>68.71</b>	<b>55.52</b>	<b>69.62</b>	<b>45.12</b>	<b>63.98</b>	<b>52.46</b>	<b>68.42</b>	+5.28	+4.58
ResNet-32	0.46	46.95	62.93	53.35	66.61	-	-	-	-	-	-	-
	ResAC-B	0.44( $\downarrow$ 4%)	<b>51.31</b>	<b>66.39</b>	<b>54.89</b>	<b>67.73</b>	-	-	-	-	+4.36	+3.46
ResNet-18	11.17	45.11	62.06	56.59	68.89	46.06	62.96	55.04	68.60	-	-	-
	ResAC-A	8.63( $\downarrow$ 23%)	<b>53.23</b>	<b>69.19</b>	<b>59.79</b>	<b>71.40</b>	<b>49.94</b>	<b>67.20</b>	<b>58.86</b>	<b>71.56</b>	+8.12	+7.13
ResNet-32	0.46	47.78	62.36	54.36	67.14	-	-	-	-	-	-	-
	ResAC-B	0.44( $\downarrow$ 4%)	<b>53.50</b>	<b>67.34</b>	<b>58.17</b>	<b>69.44</b>	-	-	-	-	+5.72	+4.98
ResNet-18	11.17	49.03	61.97	55.98	68.38	53.26	65.20	60.58	69.36	-	-	-
	ResAC-A	8.63( $\downarrow$ 23%)	<b>57.22</b>	<b>69.99</b>	<b>61.44</b>	<b>72.54</b>	<b>54.32</b>	<b>66.41</b>	<b>61.94</b>	<b>71.16</b>	+8.19	+8.02

Table 5: The CL performance of ArchCraft in *Class IL*. '#P' represents the number of parameters of the network used.

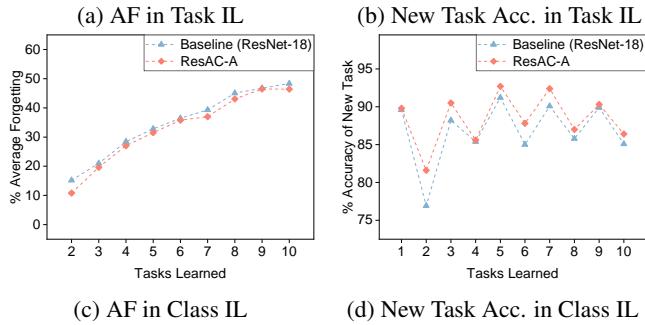
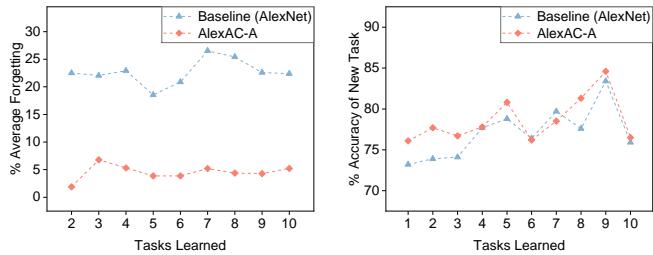


Figure 4: The average forgetting on the previous tasks and performance on the current task of the ArchCraft-guided networks and the baseline ones in *Task IL* and *Class IL*.

the similarity between these feature maps.

To put it more intuitively, we present the similarity matrix in Figure 5. Figure 5 illustrates that features extracted by the ArchCraft-guided network architecture show a higher degree of similarity across all incremental stages compared to the baseline ones. This observation suggests that the network architecture guided by ArchCraft more effectively extracts shared features between incremental tasks, facilitating adaptation to new tasks with minimal changes. Thus, ArchCraft results in better overall stability while maintaining plasticity.

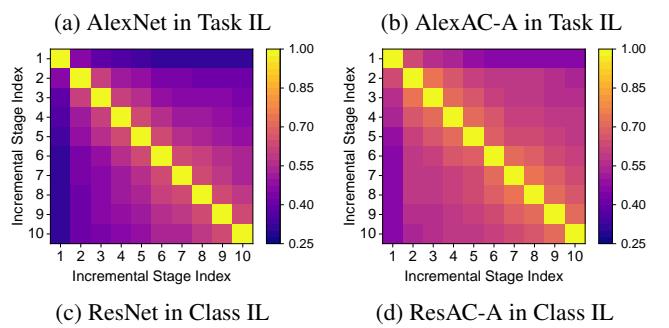
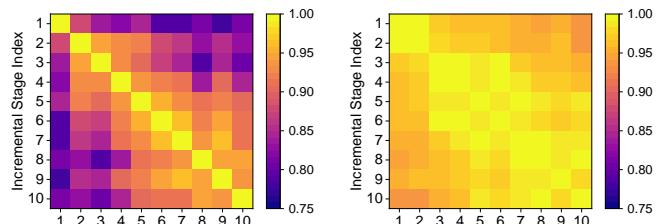


Figure 5: CKA visualization of the ArchCraft-guided networks and the baseline ones in different incremental stages across.

## 6 Conclusion

This paper identifies specific architectural designs that influence CL. The proposed ArchCraft bridges the gap between network architecture design and CL, achieving superior CL performance while utilizing a significantly more compact number of parameters than naive CL architectures. We affirm that ArchCraft holds immediate practical relevance in the design of CL-friendly networks. We hope this work inspires further study on overcoming catastrophic forgetting and enhancing CL performance from the architectural perspective.

## Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 62276175.

## References

- [Bian *et al.*, 2024] Ang Bian, Wei Li, Hangjie Yuan, Chenguang Yu, Zixiang Zhao, Mang Wang, Aojun Lu, and Tao Feng. Make continual learning stronger via c-flat, 2024.
- [Chen and Chang, 2023] Xiuwei Chen and Xiaobin Chang. Dynamic residual classifier for class incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18743–18752, 2023.
- [Elsken *et al.*, 2019] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.
- [Feng *et al.*, 2022] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9427–9436, 2022.
- [Grossberg, 2013] Stephen Grossberg. Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural networks*, 37:1–47, 2013.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [Howard *et al.*, 2019] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
- [Kang *et al.*, 2022] Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.
- [Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [Konishi *et al.*, 2023] Tatsuya Konishi, Mori Kurokawa, Chihiro Ono, Zixuan Ke, Gyuhak Kim, and Bing Liu. Parameter-level soft-masking for continual learning. *arXiv preprint arXiv:2306.14775*, 2023.
- [Kornblith *et al.*, 2019] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.
- [Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [Lee *et al.*, 2020] Janghyeon Lee, Hyeong Gwon Hong, Donggyu Joo, and Junmo Kim. Continual learning with extended kronecker-factored approximate curvature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9001–9010, 2020.
- [Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- [Lin *et al.*, 2021] Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models. *arXiv e-prints*, 2021.
- [Liu *et al.*, 2018] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.
- [Lu *et al.*, 2023] Shun Lu, Yu Hu, Longxing Yang, Zihao Sun, Jilin Mei, Jianchao Tan, and Chengru Song. Pa&da: Jointly sampling path and data for consistent nas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11940–11949, 2023.
- [Mirzadeh *et al.*, 2022a] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR, 2022.
- [Mirzadeh *et al.*, 2022b] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Timothy Nguyen, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Architecture matters in continual learning. *arXiv preprint arXiv:2202.00275*, 2022.
- [Pham *et al.*, 2022] Quang Pham, Chenghao Liu, and Steven Hoi. Continual normalization: Rethinking batch normalization for online continual learning. *ICLR*, 2022.
- [Qin *et al.*, 2021] Qi Qin, Wenpeng Hu, Han Peng, Dongyan Zhao, and Bing Liu. Bns: Building network structures dynamically for continual learning. *Advances in Neural Information Processing Systems*, 34:20608–20620, 2021.
- [Radosavovic *et al.*, 2020] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.
- [Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl:

- Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [Serra *et al.*, 2018] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR, 2018.
- [Shibata *et al.*, 2021] Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. Learning with selective forgetting. In *IJCAI*, volume 3, page 4, 2021.
- [Tan and Le, 2021] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.
- [Tiwari *et al.*, 2022] Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coresets based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2022.
- [Wan *et al.*, 2022] Xingchen Wan, Binxin Ru, Pedro M Esperança, and Zhenguo Li. On redundancy and diversity in cell-based neural architecture search. *arXiv preprint arXiv:2203.08887*, 2022.
- [Wang *et al.*, 2021] Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking architecture selection in differentiable nas. *arXiv preprint arXiv:2108.04392*, 2021.
- [Wang *et al.*, 2022] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pages 398–414. Springer, 2022.
- [Wang *et al.*, 2023] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023.
- [Yan *et al.*, 2021] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021.
- [Yuan *et al.*, 2023] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21649–21661, 2023.
- [Zenke *et al.*, 2017] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.
- [Zhao *et al.*, 2020] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13208–13217, 2020.
- [Zhao *et al.*, 2023] Haiyan Zhao, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Does continual learning equally forget all parameters? *arXiv preprint arXiv:2304.04158*, 2023.
- [Zhou *et al.*, 2022] Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv preprint arXiv:2205.13218*, 2022.
- [Zhou *et al.*, 2023] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. Pycil: A python toolbox for class-incremental learning, 2023.
- [Zhu *et al.*, 2023] Yifan Zhu, Fangpeng Cong, Dan Zhang, Wenwen Gong, Qika Lin, Wenzheng Feng, Yuxiao Dong, and Jie Tang. WinGNN: dynamic graph neural networks with random gradient aggregation window. In *The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023*. ACM, 2023.
- [Zoph *et al.*, 2018] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.